DOCUMENT RESUME

ED 354 751 FL 020 949

AUTHOR Lewkowicz, Jo
TITLE Testing Listening Comprehension: A New Approach?
REPORT NO ISSN-1015-2059
PUB DATE 91
NOTE 8p.; For the serial issue from which this paper is analyzed, see FL 020 947.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Journal Articles (080)
JOURNAL CIT Hongkong Papers in Linguistics and Language Teaching; v14 p25-31 1991

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cloze Procedure; College Students; *English (Second Language); Foreign Countries; Higher Education; *Language Proficiency; Language Research; *Language Tests; Linguistic Theory; *Listening Comprehension; Testing; Test Validity
IDENTIFIERS *China; *Summarization

ABSTRACT

Some of the issues involved in validating two English second-language listening comprehension tests are examined. The tests were two different levels of a listening summary-cloze measure developed for use in a battery of language proficiency tests in the second and fourth years of university-level study in China. In them, students listen to a talk or short lecture and complete a cloze-procedure summary of the topic, based on notes taken. The listening comprehension subtests were piloted with 203 and 117 students, respectively. Scores were correlated with those of other subtests and with total scores. Results suggest that this listening summary cloze test has a number of advantages: (1) the talk/lecture format allows the examiner considerable latitude in selection and presentation of topic, for both effectiveness and security; (2) the format lends itself to a large number of cloze items; and (3) while marking of items is largely objective, examinees are not restricted to exact words or phrases but can demonstrate their comprehension of the text. Selection of text and deletions and careful test administration are seen as crucial to test effectiveness. Further research on retention of content over time is recommended. (MSE)

HONGKONG PAPERS IN LINGUISTICS AND LANGUAGE TEACHING 14 (1991) ISSN 1015-2059

TESTING LISTENING COMPREHENSION: A NEW APPROACH?

Jo Lewkowicz
University of Hong Kong

Introduction

The aim of this paper is to discuss some of the issues involved in validating a listening comprehension test that was developed as part of a language proficiency measure for tertiary level students in China. The paper considers the need for validation, as well as some of the difficulties in validating a new subtest. It goes on to look at the rationale of the test in the Chinese context, and the theoretical basis upon which it was developed. It then describes the approach adopted, the results of piloting the test, and finally their implication for future test development.

Test Validation

Once a genuine need for a new test has been established, the test specifications have to be drawn up. These are frequently based on existing syllabi and/or needs analyses. However, as Alderson (1988) points out, both are subject to judgements being made by so-called experts, so the content of any test is dependent on the decisions made about what the learners need to know or be able to do with the language. Furthermore, test specifications have to be operationalized in a way that trivial items are avoided and items used are valid, that is, they test what they set out to test. Hence any test can only be as good as the operational realisation of its rationale, which, in reality, grows and is modified throughout the operationalization and cannot be fully developed in advance. In other words, the items of a test can only be validated against a given test rationale.

The difficulty of constructing valid items may lead to a further problem, that referred to by Davies (cited in Morrow, 1981: 19) as the reliability-validity 'tension'. To increase reliability a tester may inadvertently take a limiting view of validity which results in a reliable test of somewhat dubious, though statistically acceptable, validity.

Background

The language tester's dilemma has nowhere been greater than in China where the number of testees and the difficulty of arranging the moderation of marking inevitably lead test writers to develop objective type tests and examinations. Such tests, despite all their recognised drawbacks, are favoured because they not only encourage testers to develop a large number of test items which allow for a spread of results, but also enable comparability of results across the country. It is therefore not surprising that major tests rely heavily on multiple-choice type questions. This can clearly be seen in the two tests, that is the Matriculation English Tests (MET) and the Graded Tests for English Majors (GEM), that are being developed at the Guangzhou Institute of Foreign Languages (GIFL)--one of the major test centres in China--for widespread use throughout the country. Both tests are designed as measures of English language achievement/proficiency, the former at university entrance level, the latter at various stages throughout the degree programme.

The external constraints that were imposed on MET allowed the test developers little room for experimentation with alternative forms of testing. However, this has not been the case for GEM. The test developers were given guidelines as to the skills to be tested and suggestions were drawn up as to test format, but considerable leeway was given during initial development (1988-89) when the first two GEM examinations (GEM 4, to be administered at the end of the second year of university study, and GEM 8, to be administered at the end of the final year) were written.

The GEM examinations which are made up of four papers attempt to balance the multiple choice format with other test formats, including open-ended writing tasks. In developing the examinations a conscious effort was made to ensure a positive washback effect of the tests on classroom teaching, yet at the same time the test designers were aware of the difficulties of marking the more open-ended questions especially once the tests had been piloted and were in large scale use. It was with such considerations in mind that the new listening summary cloze test, to be discussed in this paper, was developed.

The Testing of Listening Comprehension

The testing of listening comprehension has undergone considerable changes in recent years as a result of a greater understanding of the processes involved in listening (see e.g. Brown, 1990, Buck, 1988, Rost, 1990). There has been a movement away from the belief that listening comprehension is a one-way bottom-up process towards the theory that it requires a more complex combination of top-down and bottom-up processing. In other words, listening is not just the consecutive processing of subskills starting from the lowest level to the highest. It requires processing at a number of different levels to which no fixed order can be attributed. The listener utilizes not only his knowledge of the language being spoken, that is, his knowledge of the morphemes, phonemes. lexis and syntax of the language, but also his knowledge of the outside world and how it relates to the topic at hand, as well as his interpretation of what has been said so far, in order to comprehend the message. The listener hypothesizes about what the speaker will say next and uses his knowledge of the culture to help him understand linguistic complexities. He will therefore rely heavily on context and it is with regard to context that tests of listening comprehension have changed dramatically. There is now a recognition of the fact that "if we ask students to decode short decontextualized sentences, we are not testing listening comprehension at all but asking students to engage in a very unnatural activity which seems to be confined largely to the second language classroom". (Buck, 1988: 22). As a result, it is becoming less and less common to see decontextualized utterances being used in tests of listening comprehension.

What appears not to have changed is the belief that testers should strive towards designing tests that are 'pure'. Tests that require students to respond in speech or writing are considered 'contaminated' because they are testing more than the skills of listening. Buck (1988: 33), for example, cautions against mistaking intervening variables and suggests ways to overcome the 'problems' of contamination. Yet if we look at the skill of listening as it is used in real-life, it will soon become clear that we rarely listen solely for pleasure or enjoyment. More often than not the listener is required to 'do' something with what he has heard: to respond to it in speech, take notes on it, or reproduce it in written form. Researchers are beginning to assert the fact that students should not only be tested on their level of comprehension, but also on their ability to use the information they have heard, even though they recognise that such more integrative tests may affect the reliability of the measure.

> Although judgements on the value of listening-skill tests primarily on the basis of statistical reliability are desirable on scientific grounds, such judgements are questionable in terms of educational principles....The unfortunate effect [of such tests] on the pedagogy (which is not necessarily intended by the writers) is that the test users--language teachers and learners--may come to accept that listening ability amounts to whatever can be measured reliably... (Rost, 1990: 180).

Testers, therefore, should be willing to sacrifice test purity in order to maximize task authenticity and provide a positive washback effect on teaching.

Listening Summary Cloze

One of the skills required of tertiary level students is the ability to understand and retain information imparted during lectures. The most commonly used means to achieve this end is through notetaking. Yet good notetaking in a second or foreign language is not an easy task. It may, in certain situations, be an important skill to teach and hence to test. Students studying English in China, for example, may be required to attend lectures on literature and culture given by native speakers of English or may want to go on to study abroad. They would benefit from notetaking being included as part of their syllabus. A similar argument would hold true for students studying in an L2 at tertiary level, as in Hong Kong. If the teaching is to be effective, then the skill needs to be tested. However, the testing of notetaking is not easy. Notetaking is a very individual skill which cannot be realistically assessed on its own. It can, however, be tested indirectly by asking students to take notes on a lecture but judging them on the accomplishment of a task based on their notes.

A number of such tasks are in widespread use, varying from the more closed, objective tasks of completing an outline of a talk/lecture, to the more open, subjective tasks such as writing a summary of a lecture. The problem with the former is that it is not only unrealistic, but allows for only a limited number of test items. The problem with the latter, on the other hand, is the difficulty of the task as well as the marking. A desirable test would therefore be one that minimized these disadvantages, but at the same time had a positive washback effect on teaching.

To meet these criteria the listening summary cloze test (LSC) was introduced as part of the GEM examinations. This test is made up of two tasks:

1. listening to a talk/lecturette and taking notes on it;
2. completing a summary cloze passage based on the notes taken during the listening and notetaking phase.

The first phase of the test, which is not assessed, is designed to facilitate completion of the second phase, and at the same time to encourage students to take notes while listening. The students use these notes to complete the second phase of the test, a summary cloze passage for which a rational deletion procedure is used. Deletions are placed on high information content to avoid students being able to complete the summary without having heard the talk/lecturette. One mark is awarded for each correct word or phrase used to complete the blanks.

The LSC test described above is similar to the listening cloze (LC) test described in Buck (1988: 28-29) in that in both tests the cloze passage is based on a summary of the original text. However, in the LC test the students are given the summary prior to listening to the talk/lecturette and they can complete it while listening or at the end from memory, whereas in the LSC the summary is withheld until the testees have heard the text and completed their notes. Therefore, they have to understand the talk/lecturette to be able to take notes on it, and then they have to rely on these notes to complete the summary.

Development of the Test

Two 20-item listening summary cloze test were developed in 1989, one as part of GEM 4 and the other as part of GEM 8. The former was piloted with a total of 203 students--173 end of first year English majors and 107 end of second year English majors at GIFL. The latter was also piloted at GIFL, on 117 end-of-third-year English majors. There was some delay in piloting the tests and as a result fourth year students were not targeted during the first trial of the test.

The lecturettes used in GEM 4 and GEM 8 were on different topics. An additional difference was the order in which the information was given in the summary. In GEM 4, the summary followed the order of the original text, while in GEM 8 the order of the summary was different to that of the original text, presumably making it more difficult for the students to retrieve the relevant information from their notes.

Results of Piloting

After scoring the tests, the results were analyzed using SPSS. The results of the listening summary cloze were correlated with the results on the other subtests, including the rest of the listening subtest, as well as the total scores on the test battery. In addition, facility values and discrimination indices were calculated for each item. A summary of results is given in tables 1-4 below.

Table 1: Mean and Standard Deviation of the Listening Summary Cloze

	Items	Mean (%)	Std Dev
GEM 4 (Yr 1)	20	8.3584 (42.8)	2.7637
GEM 4 (Yr 2)	20	11.4112 (57.1)	2.8449
GEM 8 (Yr 3)	20	9.7094 (48.5)	3.6861

Table 2: Correlation of Listening Summary Cloze (LSC) with a 30-item Listening Subtest and Total Test Scores

	Listening	Total Test Scores
LSC (Yr 1)	.3865**	.7478**
LSC (Yr 2)	.4644**	.6783**
LSC (Yr 3)	.5679**	.6650**

1-tailed significance ** .001

Table 3: Summary of Item Analysis for Listening Summary Cloze (GEM 4)

	Year 1		Year 2	
Difficulty	Total	No (<.3)	Total	No (<.3)
Very Difficult	1	0	2	1
Difficult	6	0	3	0
Intermediate	6	1	6	0
Easy	4	0	5	0
Very Easy	1	0	4	3

Table 4: Summary of Item Analysis for Listening Summary Cloze (GEM 8)

Difficulty	Total	No (<.3) .
Very Difficult	1	0
Difficult	5	0
Intermediate	10	0
Easy	4	1
Very Easy	0	0

The mean scores indicate that the tests were of moderate difficulty. Not surprisingly, the second year students performed better on the test than the first years. A comparison of means across subtests is impossible since, as mentioned above, the two were quite different.

The correlations with the rest of the listening test suggest that though there is significant overlap between the two, the listening summary cloze tests skills other than listening. In terms of test purity these tests can indeed be seen as contaminated, which was to be expected because of the dual nature of the tests. However, for both the second and third year results the correlations of the listening summary cloze with the listening subtest were higher than for the other subtests suggesting that the listening summary cloze was functioning to a large extent as a listening test and not one of reading and filling in the blanks. The exception was with the first year results where the correlation between reading and the listening summary cloze was considerably higher than that for listening and listening summary cloze, the correlations being .6262** and .3865** respectively. This may have been because the first years are less proficient in the skills being tested since it is only once they start their degree programme that listening skills are intensively taught and hence they may still have to rely on reading skills to complete the test. This result was not unduly disturbing as the test was designed to be administered at the end of the second year and was somewhat difficult for the first year students.

The item analysis is perhaps most interesting. Both listening summary cloze tests appear to have needed only limited moderation. For GEM 4 a maximum of 5 items needed to be looked at and possibly changed or omitted, whereas for GEM 8 only two items were unacceptable. Such a modest amount of moderation, notwithstanding Fletcher's claim that "in future the test needed to be constructed with more care" (1990: 66), was to be expected.

On the basis of these results it was decided to retain the summary listening cloze as part of both the GEM 4 and GEM 8 battery of tests. The same test format has in the last two years (1989-91) been experimented with at the Language Centre of the University of Hong Kong with equally encouraging results.

Discussion

The listening summary cloze described in this paper appears to have a number of advantages. The talk/lecturette format allows the setter considerable flexibility in the choice of topic and the degree to which the topic is developed. The test may be made easier or more difficult by the length of the lecturette, the topic chosen or the method used by the speaker(s) to develop the topic. The lecturette can be purpose written and recorded as was the case at GIFL, or prerecorded from, for example, the radio, as is the practice at the University of Hong Kong. The latter is more authentic but depends on a bank of material being built up for possible future use for testing.

Another distinct advantage of this test format is that it lends itself to a large number of items being written. This not only eases the pressure at moderation since some of the items can always be eliminated if they do not work, but it also ensures a spread of marks. The validity of the test depends on the validity of the texts chosen and the items deleted. In order to ensure that both lower and higher-order listening skills are tested, the rational deletion procedure is used. This allows the test setter to make judgements as to the most suitable items for testing and enables a balance of listening enabling skills to be sampled. However, it must be remembered that the testing of lower-order skills is easier and unless due care is taken to construct a balanced test, test validity will be sacrificed.

The marking of the test items is largely objective, yet the testees are not restricted to the exact words or phrases used on the tape. This allows them to demonstrate their comprehension of the text and adds to the authenticity of the task. Furthermore, to complete the task the students need their notes; they therefore have a purpose for taking notes and they need not rely on memory for task completion. To test that the students can really use their notes, the setter can vary the order of information given in the summary. It is likely that following the exact order of the lecturette makes the task easier, but often in real life one has to be not only selective in the information to be used, but also aware that the order in which information is given is not sacrosanct.

Like all objective type tests, the listening summary cloze is not easy to set. Setters have to be careful in choosing the listening text to ensure it lends itself to a good summary and that there is sufficient redundancy of information to allow for notetaking. They must also pay particular attention to the deletions to ensure that these cannot be filled in on the basis of general knowledge, on the one hand, and that they are not testing trivial information likely to be missed by the listener, on the other. Thus pretesting on colleagues who have not heard the tape is advisable, and the moderation of the marking scheme is vital.

Careful administration of the test is also crucial. Clear instructions to the testees are essential, as they are for any test, and guidance as to the nature of the information to be noted may also be necessary. There are numerous ways testees can be prevented from filling in the summary cloze while listening. The summary can be withheld until after the tape has been played or the test booklet can be prepared in such a way that the students are prevented from looking at the summary while listening. Whatever method is used, it is essential that the students do not fill in the gaps while listening as they are likely to miss essential information.

Conclusion

The validation of this test method is still in its early stages and more research needs to be undertaken. However, the results to date are encouraging and indicate the suitability of the test for use at tertiary level. When the GEM examinations go nationwide in 1991-92 it is hoped that this test will have a positive washback effect on teaching and will encourage staff in China to move away from the more traditional approaches of teaching listening to more integrative ones which allow a greater variety of activity in the language classroom. It should also help students see the merit of good notetaking.

The test itself, though an indirect measure of notetaking and listening, does appear to have face validity. It would be interesting to see if students could fulfil the task of completing the summary cloze after a delayed period, such as a week, as this would verify that the students really are capable of using the notes they have taken during a lecture.

A considerable effort has to be made when writing the tests to ensure that there are no ambiguities in the summary and that the test is tapping relevant notetaking/listening skills. It would therefore seem advisable, even in a country as large as China where the problems of test security are well known, to build up a bank of tests and not to rely on having a new test for each successive administration. This would not only be a more cost effective approach, but also a more theoretically sound one.

Finally a word of caution. There is no ideal test and the listening summary cloze is no exception. Although it is a test that appears to satisfy many prerequisites of a good test and therefore may well be suitable for advanced learners needing notetaking and listening skills, it should be considered in conjunction with other test formats and not as a substitute for all other listening tests.

Acknowledgements

First and foremost I would like to thank Wang Chuming who was Acting Head of the Testing Project at GIFL when this test was developed. It was thanks to him that this test was piloted and that I received the results discussed in this paper. I would also like to thank the rest of the 1988-90 testing team at GIFL for their help in developing and moderating the test. Finally, I would like to thank Peter Falvey and Desmond Allison for their invaluable comments on an earlier draft of this paper.

REFERENCES

Alderson, J. Charles 1988. 'New Procedures for validating proficiency tests in ESP? Theory and Practice'. *Language Testing* 5/2: 220-232.

Brown, Gillian 1990. *Listening to Spoken English.* Longman: London and New York.

Buck, Gary 1988. 'Testing Listening Comprehension in Japanese University Entrance Examinations'. *Japanese Association of Language Teachers Journal* 10/1 & 2: 15-42.

Fletcher, Nicholas 1990. 'Test Validation at G.I.F.L.--Towards a More Scientific Approach'. In Chen Lie Qiang, Chen Xin Seng & Zhang Mei Qin (eds.) *The Proceedings: Second Guangdong Symposium on English Teaching of Foreign Experts/Teachers.* May 25 1990: 59-70.

Morrow, Keith 1981. 'Communicative Language Testing: Revolution or Evolution?' In Alderson, J. C. and Hughes, A. (eds.) *Issues in Language Testing, ELT Documents 111* pp. 9-26. The British Council.

Rost, Michael 1990. *Listening in Language Learning.* Longman: London and New York.